SUBSTTITUE SPECIFICATION

Method for capturing utilization behavior of an Internet/Intranet subscriber by combining a plurality of methods

BACKGROUND OF THE INVENTION

Subject of the Invention

The subject of the patent application (subsequently referred to as the invention) concerns a method for capturing the utilization behavior of an Internet/Intranet subscriber, including the features of claim 1.

DESCRIPTION OF THE RELATED ART

Unauthorized utilization of Internet access (e.g. by private individuals using the flat-rate tariffs of an Internet Service Provider or by company employees) can reach levels that pose a threat to the business viability of an Internet Service Provider or a company. In technical terms, it is currently very difficult to identify such utilization behavior early enough to respond at the appropriate time.

Private customers, business customers and company employees use the Internet/Intranet in different ways, e.g. for occasional surfing, for data throughput (variable quantities), for playing games, etc. Until now, it has been very difficult for the Internet Service Provider to implement selective marketing activities for different user groups, to identify

market trends, or even to determine the potential for cost savings within a company (e.g. in the case of corporate networks), because the technical tools available cannot satisfactorily associate individual Internet/Intranet users with different behavior categories with accuracy.

Various tools currently exist to detect fraud in the context of telecommunications. These tools are based on a variety of techniques, such as the rule-based approach or neural networks, etc. These techniques are used to evaluate CDR (Call Detail Records) or signaling data from the CCS7 signaling system.

A tool called "HP OPENVIEW SMART INTERNET SUITE; SMART INTERNET USAGE" has been announced. This tool collects, correlates and compresses utilization-specific Internet data, and offers a retrieval function for this data (data mining). Details of the technical implementation and exact scope of the retrieval function are not known.

SUMMARY OF THE INVENTION

The subject matter of the application addresses the problem of providing a method, which increases the significance of information output and reduces error rates, in comparison with conventional methods.

The problem is resolved using a method with the features of claim 1.

As well as the early collection of suspicious incidents (keyword: fraud) regarding non-specific utilization (e.g. unauthorized utilization of the Internet/Intranet), this tool produces results that can be used for marketing purposes, to identify market trends, to ensure rapid response to requirements to expand the Internet/Intranet network, or to reduce the cost of Internet/Intranet utilization.

The invention gives Internet Service Providers and companies extremely good information about the type of Internet/Intranet utilization (and particularly unauthorized utilization), market trends (and particularly sudden behavior changes in relation to Internet/Intranet utilization), marketing, and the necessity for network expansion etc. In particular, this invention overcomes the disadvantages of the data mining tools that are currently on the market. By configuring the subject matter of the application in a particular way, the results of individual methods can be compressed (combined and associated) to provide significant information with extremely low error rates.

Advantageous developments of the subject matter of the application are specified in the subclaims.

BRIEF DESCRIPTION OF THE DRAWINGS

In order to allow an adequate understanding of the subject matter of the application, the following implementation example provides an explanation on the basis of figures, in which

Fig. 1 shows a known sequence;

Fig. 2 shows a schematic block diagram of elements and their interaction in the subject matter of the application;

Fig. 3 shows an application example for modeling a behavior category in the causal network.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

The same designations in the figures refer to the same elements.

As shown in Figure 1, the known tool appears to work as follows: In Step 1, the Internet data Idat is correlated and compressed in accordance with fixed rules as part of preprocessing (PP), to produce Internet Data Records (IDR); in Step 2, the Internet Data Records can be output together with rules entered by the operator (e.g. in the form of select statements) as the OUT result as part of the analysis procedure RETR (retrieval, data mining). A number of disadvantages are apparent as follows: A change in the IDRs requires a change in the preprocessor; the rules can relate only to the IDRs (and

particularly the IDR structure); the IDRs are static, which means that information is lost; purely rule-based systems are not capable of learning; purely rule-based systems do not recognize "exceptions to the rule" (resulting in a large volume of incorrect information); and purely rule-based systems do not recognize limit ranges.

Using the method as per the invention shown in Figure 2, the Internet data Idat can undergo rule-based preprocessing (RBPP), where the Internet data is correlated and compressed. The Internet data can be stored as an interim result in interim memory (INTM), in preprocessed format if applicable. The Internet data, in preprocessed format if applicable, and after interim storage if applicable, is then transferred to a method approach (MA) unit, which has a rule-based approach (RBA), a neural network with supervised training (NNUE), density-based profile modeling (DBPM), and a causal neural network (KNN), subsequently referred to as a causal network. As shown by two two-directional arrows, the method approach (MA) unit works with a rule base (RB), which contains rules, an MO/TR database, which contains modeling and training data, and a HIST database, which contains the results of analyses from current and past monitoring periods. The interim results output by the method approach (MA) unit or stored in the HIST database can undergo

analysis in the combination (COMB) unit and then be output as result OUT.

The method as per the invention includes a combination of four different method approaches, namely the rule-based approach and three neurocomputing method approaches (neural network with supervised training, density-based network and causal network). Both the formulation of rules and modeling with neurocomputing methods are based on data that is stored by the Internet Service Provider or company: RADIUS Accounting data (generally stored), TCP dump protocol data (stored if required, variable volume), SNMP (Simple Network Management Protocol) data (stored if necessary), etc. The resulting model represents a controlling and marketing tool.

The method as per the invention is intended for use with Internet/Intranet data that is stored by the Internet Service Provider or a company. Such data includes RADIUS Accounting data, TCP dump data, and SNMP data. The method can also process all other types of Internet/Intranet data.

RADIUS Accounting data is derived from data as described in IETF Specification RFC 2139. An actual implementation example is described in Livingston Enterprises Inc., Radius dictionary, V1.6, 1997.

TCP dump data is derived from data as described in the UNIX man pages 'tcpdump - dump traffic on a network'.

SNMP data is derived from data as described in the various IETF RFCs. An actual implementation example is described in Livingston Enterprises Inc., Configuring SNMP, Manual Portmaster 3.

An optional rule-based preprocessor can be installed to accelerate the processing of data. The task of the preprocessor is to correlate and compress the Internet/Intranet data, in order to provide data records with the attribute values required by the actual method.

In principle, a preprocessor can be used in the same way as is intended in the known solution proposal. However, this is conditional upon the IDRs containing a superset of the attribute values required by the method.

A rule-based preprocessor is used in a preferred embodiment of the invention. Using this configuration, the rules control the correlation and compression of Internet/Intranet data.

If a new characteristic attribute is added or an existing characteristic attribute is omitted in the actual method, then the selection rules of the preprocessor can easily be adapted (automatically). Automatic adaptation of the selection rules can be controlled by means of notifications (unsolicited messages) to the preprocessor, as shown by ADAP (adaptation) in Figure 2.

The method described below can be based on any of the following:

- on the Internet/Intranet data directly
- on results from any preprocessor (e.g. HP (Hewlett Packard) IPR)
- on results from a specific rule-based preprocessor (shown by INTM in Figure 2).

The actual method is divided into four methods. Each method uses a different method approach. The four different method approaches are as follows:

- the rule-based method
- the neural network - supervised learning
- density-based profile modeling
- the causal network.

With the rule-based method, typical, user-specific behavior categories can be modeled with the aid of rules. Behavior is then classified by a behavior category. For example, behavior categories such as "Student private use", "Employee private use", "Self-employed private use", "Small business establishment use", "Large business establishment use", "Games player", "Internet/Intranet addict", "User with high mail volumes", etc. can be expressed in the form of rules using their own characteristic properties. The rules are applied to all Internet/Intranet data or part of this data (e.g. the result of

preprocessing). As a result of the method, each user can be assigned to no behavior categories, one behavior category, or several behavior categories, at the end of a monitoring period the monitoring period can also vary depending on the behavior category and the reason for monitoring. For example, if the purpose is to detect fraud (as part of controlling), then the monitoring period would be quite short (e.g. t=1 day). However, a monitoring period of several weeks would be used to obtain marketing information (e.g. t=4 weeks). If the results of each monitoring period t(i) are stored in the HIST database on a user-specific basis, then it is very easy to identify changes in user behavior by comparing the individual results for t(i). For example, the utilization of a given user may have matched to the behavior category "Student private use" at the beginning, but now it would be more appropriate to assign it to the behavior category "Small business establishment use".

The objective is to formulate rules for each behavior category. These rules are defined by means of logical expressions, where the fields (attributes) of the different data records are used as variables. For example:

"Private contract employee" utilization ::= the following applies to all data records in the monitoring period: utilization time Monday to Friday between 17:00 hours and 24:00 hours and utilization time on weekends from 00:00 hours to 24:00

hours and data throughput rate < 2 megabytes per utilization and maximum utilization time = 2 hours.

"Private contract self-employed" utilization ::= there is one data record in the monitoring period, for which the following applies: not "Private contract employee" utilization and the following applies to all data records in the monitoring period: data throughput rate < 10 megabytes per day and maximum utilization time = 8 hours.

In principle, rules can refer to one or more data records (including different files).

During the usage phase of the rule-based method, all the selected rules are tested for the specified data at the time t. The results are initially recorded on a user-specific basis in the HIST database.

Strengths of the rule-based method:

- Classification of user behavior in the form of behavior categories

- It is easy to derive trends, marketing information, contract infringements, etc.

With the supervised approach, a neural network is trained with a set of examples. The requirement for training is that the corresponding target value must be specified for each example, i.e. it must be known at the time of training whether e.g. a defined use existed or not for the example concerned

(examples of defined uses are "Contract infringement private contract employee", "Primary use surfing", "Primary use games player", etc.). It is also necessary to specify the target values to be examined and the characteristic attributes for the example. The characteristic attributes determine the behavior of a user. The behavior is therefore dependent on specific attribute values (the data itself).

Examples of characteristic attributes are as follows:

- average utilization time per day over a monitoring period (e.g. four weeks) for the user
- distribution of utilization time for the user
- maximum utilization time
- minimum utilization time
- average throughput rate per day over a monitoring period (e.g. four weeks) for the user
- distribution of throughput rate for the user
- maximum rate
- minimum rate
- average utilization duration of special Internet/Intranet services over a monitoring period (e.g. four weeks) for the user
- distribution of utilization duration for the user
- maximum utilization duration
- minimum utilization duration

etc.

During the training phase (preliminary stages) of the neural network, the objective is to create a model that can decide, based on the specified example, whether or not the Internet/Intranet access utilization relates to one or more of the defined target values, for a given user. The model is created by means of supervised training, the principles of which are described in detail in 'Learning internal representation by error backpropagation' by D.E. Rumelhart, G.E. Hinton and R.J. Williams, contained in 'Parallel Distributed Processing', Pages 318-362, Cambridge, MA, MIT Press (1986).

The following steps are performed during the training phase: A behavior sample is allocated to each user in the form of attributes. This behavior sample describes a specific profile over an extended period. In this context, the attributes characterize utilization relating to a defined target value. The period that is used as a basis for the behavior sample should not be shorter than four weeks, and should precede the time when the method is used for the purpose specified above.

Based on training data, the neural network is trained for utilization relating to the defined target values. The training data indicates whether or not the utilization can be assigned to a specific target value.

During the usage phase of the neural network, which begins when the training phase is complete, the following steps are performed continuously:

Using the examples as a basis, the neural network decides whether or not the utilization can be assigned to a specific target value. This decision is recorded on a user-specific basis in the HIST database, as a result of the monitoring period.

If necessary, the neural network can be trained with new target values relating to its utilization (e.g. as yet unknown cases of contract infringement).

These methods are applicable if the user is included in the data.

Strengths of this method:

- Simple assignment of exceptions;
- Exceptions included in the result;
- Learning capability.

Density-based profile modeling is concerned with the probabilistic modeling of the behavior for each user (probabilistic profile modeling), i.e. a model is generated for each user, based on the examples associated with that user. These examples comprise characteristic attributes and specific attribute values, which describe the use of the Internet/Intranet relating to one or more target values.

Examples of characteristic attributes are given in the previous section.

The following steps are performed during the training phase of density-based profile modeling: Each user is assigned a set of examples, which describes the behavior of the user over an extended period. The period that is used as a basis for the behavior sample should not be shorter than four weeks, and should precede the time when the method is used for control and marketing purposes. A probabilistic profile is generated for each user. This is produced by means of density estimates using the EM algorithm. The exact description is given in 'Neural Networks in Pattern Recognition' by Chris Bishop, Oxford Press (1996). Once the training phase is complete, the usage phase of density-based profile modeling can begin, during which the following steps are performed continuously:

The data for e.g. one day is analyzed with regard to the data contents specified for probabilistic profile modeling (a new example is produced). The density-based profile model outputs a value that represents a probability for the Internet/Intranet utilization of the observed entity with regard to the defined target values. This value is recorded. If this value differs from the previous values beyond a specified threshold value, then a recommendation is given that the result should be indicated in each case. Using this method, it is very

easy to detect sudden changes in Internet/Intranet utilization. The profile model is retrospectively adapted using the current example. This method can be used if the user is included in the data. Strengths of this method: Detection of sudden changes in user behavior; learning capability.

The causal network method is based on the modeling of typical behavior scenarios in the form of causal dependencies and probabilities of specific data contents as in the example shown in Figure 3: "Private use employee". A Private use employee (PA) is assigned a specific utilization time (UC for UseClock), a specific utilization duration (UT for UseTime), and a specific throughput rate (RATE). The days of the week affect the level of utilization time, utilization duration and throughput rate, depending on whether it is a weekday (WD) or a weekend (WE). The causal dependencies are based on the analysis of known cases. They do not have to be assigned to specific users. The following steps are performed during the modeling phase of the causal network: the causal dependencies with regard to data contents are formulated for all data. Appropriate probabilities are assigned at those places where causal dependencies exist. The domain knowledge of technical experts is required during the modeling phase. The principle of the causal network is described in 'An Introduction to Bayesian Networks' by Finn V. Jensen, UCL Press (1996). The following steps are

performed continuously during the usage phase of the causal network: The data records of the data to be examined are continuously examined for the causal dependencies that have been formulated. For each user or event, the system decides the probability of a specific use occurring with regard to the defined scenario. This decision is recorded in the HIST database as a result of the monitoring period. If the user is included in the data, then the results are recorded on a user-specific basis. The probabilities behind the causal dependencies can be retrospectively adapted. The causal dependencies of new, previously unknown categories are added to the existing causal dependencies if required. This method can also be applied if the user is not included in the data. However, it is not possible to assign a category to a specific user in this case. Strengths of the causal network method: Assignment of limit ranges in behavior categories; recognition of limit ranges in behavior categories; learning capability. In principle, it is possible to output the individual results of the individual methods. In a further form of the invention, the individual results of the individual methods are compressed into an overall result. This compression includes the individual results of the various methods. The individual results can be taken from the current monitoring period and from previous monitoring periods. Each of the described methods has specific strengths. This can be useful

for compression, as described in the following example: The rule-based method may unequivocally assign User x to the "Private use self-employed" behavior category, if he used the Internet for more than 2 hours on a single day (based on the example rule described above). However, the causal network may consider User x falls into the "Private use employee" behavior scenario, since he observed the utilization time of less than 2 hours in more than 90% of data records, for example. These findings could then be shown in the overall result, in such a way that utilization of the Internet/Intranet by User x is treated as "Private use employee" with a few minor exceptions. Another example of using compression is to identify trends by analyzing the results of different monitoring periods.